ASLA-information 24:1. 1998. 33–40.

Lars Borin
Uppsala universitet
Institutionen för lingvistik
Box 527
751 20 Uppsala
Lars.Borin@ling.uu.se

ETAP: Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter

Bakgrund

ETAP (*Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter*) är ett projekt inom forskningsprogrammet *Översättning och tolkning som språk- och kulturmöte*, som finansieras av Riksbankens jubileumsfond och genomförs i samarbete mellan de språkvetenskapliga sektionerna vid universiteten i Stockholm och Uppsala, och som planeras löpa under de sex åren 1996–2001.

Arbetet inom ETAP-projektet utförs vid Institutionen för lingvistik, Uppsala universitet.

Medarbetare

I projektet medarbetar eller har medarbetat:

Professor Anna Sågvall Hein (projektledare 1996–97;
medarbetare fr.o.m. 1998)
Universitetslektor Lars Borin (medarbetare 1996–97;

projektledare 1998–99)
Universitetslektor Erik Tjong Kim Sang (medarbetare 1996–)
Universitetslektor Hong-Liang Qiao (medarbetare januari–juni 1996)
Forskarstuderande Klas Prütz (medarbetare 1996–)
Projektassistent Jörg Tiedemann (medarbetare 1997)
Förste forskningsingenjör Bengt Dahlqvist (medarbetare 1996–)
Forskningsingenjör Per Starbäck (medarbetare 1996–97)
Studenter på Språkteknologiprogrammet (det 160-poängs magisterprogram som institutionen ansvarar för)

Samarbetspartners utanför projektet

PLUG-projektet
Scania AB i Södertälje
Stiftelsen Invandrartidningen i Stockholm

Mål

Projektets mål kan delas upp i fyra delmål, där arbetet delvis har bedrivits parallellt:

(1) Upprättande av en flerspråkig parallellkorpus, d.v.s. en textkorpus där de ingående texterna finns i parallella versioner på mer än ett språk. I praktiken rör det sig här alltid om översättningar och i ETAP-korpusen är det översättningar med svenska som källspråk till ett antal olika målspråk. Upprättandet innefattar insamling, i förkommande fall inläsning med

scanner, formatkonvertering och SGML-uppmärkning av textmaterialet. Korpussammansättningen beskrivs i nästa avsnitt.

(2) Länkning av parallelltexterna ner till meningsnivå (d.v.s. ange vilken mening i texten på språk B som är översättning av en viss mening i motsvarande text på språk A, och så vidare, för var och en av meningarna i de två texterna). I första hand avses här länkning mellan svenska och vart och ett av de andra språken, på grund av det sakförhållandet att svenska är källspråk och de andra texterna var för sig utgör översättningar av den svenska texten.

(3) Ordklasstaggning av texterna

(4) Utveckling av metoder för att länka texterna på lägre nivåer än meningsnivå, t.ex. fras-, ord- och morfemnivå, för att hitta översättningsekvivalenter, eller flerspråkig lexikal information. Åter koncentreras arbetet på de språkpar där den svenska källtexten utgör den ena medlemmen.

Dessutom är ett viktigt övergripande mål för projektet att de resurser, i form av korpusmaterial och datorlingvistiska verktyg, som utvecklas inom dess ram är modulära och återanvändbara — eftersom de kostar så pass mycket att utveckla, räknat både i pengar och i mänsklig möda — så att de kan användas både inom andra av översättningsprogrammets projekt och inom annan forskning vid institutionen, både idag och i framtiden. Av samma skäl bedrivs också ETAP-projektet och PLUG-projektet[1)] i nära samarbete.

Korpus: Textmaterial, språk, m.m.

I den färdiga korpusen skall ingå tre textmaterial:

(1) Teknisk dokumentation (verkstadshandböcker) från Scania i Södertälje, i parallella versioner på svenska, engelska, finska, franska, italienska, nederländska, spanska och tyska.

(2) Ett antal nummer av Invandrartidningen från år 1997, i parallella versioner på svenska, engelska, finska, polska, serbokroatiska och spanska.

(3) Den svenska regeringsförklaringen fr.o.m. 1988, i parallella versioner på svenska, engelska, franska, spanska och tyska.

Med undantag för Invandrartidningen där materialet tidigare har varit svårt att få fram, men där nu (maj 1998) inläsning med scanner och korrekturläsning pågår, är de första två delmålen uppfyllda (se Tjong Kim Sang 1996a, 1996b). Projektet arbetar således för närvarande huvudsakligen på delmål nummer tre och fyra. Dock pågår också arbete i samarbete med PLUG-projektet med att utveckla en databas och databasåtkomstverktyg för flerspråkiga korpusmaterial.

Ordklasstaggning

När det gäller ordklasstaggning har vi valt att använda oss av Brilltaggning, en icke-statistisk metod, som har tränats på svenska och franska textmaterial (Prütz 1997). När det gäller taggning av projektets andra språk, undersöker vi för närvarande olika möjligheter. Ordklasstaggare, vare sig de är statistiska eller av annat slag, behöver i allmänhet läras upp, vilket man gör med korrekt taggad träningstext, något som kan vara svårt att få fram för flera av de språk vi arbetar med. Det

existerar dock också metoder för att träna taggare direkt med hjälp av otaggad text, men dessa metoder, liksom statistiska metoder i allmänhet (s.k. ”kunskapsfattiga” metoder) kräver betydligt större mängder text än vad som står till vårt förfogande. Därför utforskar vi nu framför allt sådana metoder som innebär att man drar nytta av — eventuellt i kombination med statistiskt baserade metoder — den omfattande lingvistiska kunskap vi redan har om källspråkstexten, antingen därför att den är ordklasstaggad eller därför att den har fått en fullständig syntaktisk analys med hjälp av den syntaktiska parser (UCP) och grammatik för svenska som utvecklats vid institutionen och som används i andra datorlingvistiska projekt vid institutionen[2)]. Ordklasstaggning av målspråkstexterna och länkning på ordnivå mellan käll- och målspråkstexter blir ur detta perspektiv samma, eller åtminstone närbesläktade problem.

En fråga som vi behöver ta ställning till i detta sammanhang är den om vilken tagguppsättning eller vilka tagguppsättningar som skall användas för de olika språken, där vi ställs inför de delvis motstridiga kraven att tagguppsättningen för ett visst språk bör vara maximalt trogen det språkets struktur, och att länkning mellan språken rimligen underlättas av att man använder en gemensam tagguppsättning för samtliga språk.

Utvinning av flerspråkig lexikal information: länkning på ordnivå

Länkning på ordnivå är det viktiga första steget då man använder sig av parallellkorpora för att utvinna flerspråkig lexikal information, som sedan kan användas exempelvis i en applikation som maskinöversättning eller översättningsstöd (Sågvall Hein 1997). Det man egentligen är ute efter är

naturligtvis länkning på *lexikal* nivå, om vi i god lexikalistisk anda såsom fallande inom lexikonets domvärjo räknar såväl delar av ord (morfem) som olika slags flerordsenheter (fraser), och även en del syntaktiska mönster. Det som är lätt att urskilja i de språk som ingår i vår korpus är dock (graf)ord, och därför är det lämpligt att börja med dessa[3)].

Vi har hittills genomfört en empirisk undersökning av ett antal olika ordlänkningsmetoder, både statistiskt baserade och andra, på Scaniadelen av korpusmaterialet (Tiedemann 1997, 1998). Metoder för att kvantifiera ordlikhet har också undersökts med avseende på hur den mängd lingvistisk information de använder sig av påverkar deras förmåga att urskilja potentiella översättningsekvivalenter i parallelltexter ur Invandrartidnings-materialet (Borin 1998).

Noter

1) PLUG, som skall uttydas *Parallellkorpora i Linköping, Uppsala och Göteborg*, är ett samarbetsprojekt mellan Institutionen för datavetenskap vid Linköpings universitet, Institutionen för lingvistik vid Uppsala universitet och Institutionen för svenska språket vid Göteborgs universitet. PLUG-projektet finansieras av HSFR och NUTEK inom ramarna för det andra Språkteknologiprogrammet. Det syftar till att utveckla och utvärdera gemensamma resurser för hantering, annotering, lagring, länkning och lingvistisk informationsutvinning ur flerspråkiga parallellkorpora. Uppsalas arbete inom PLUG-projektet leds av professor Anna Sågvall Hein.

2) Dessa är: SCARRIE-projektet, ett EU-projekt för att utveckla korrekturläsningsstöd för skandinaviska språk, med deltagare från universitet, mjukvaruindustri, tidnings- och bokförlag i Sverige, Norge och Danmark, samt ett samarbetsprojekt med Scania AB i Södertälje om utveckling av språkgranskningsverktyg och översättningsstöd för deras produktion av teknisk dokumentation. För båda projektens gäller att arbetet vid vår institution leds av professor Anna Sågvall Hein.

3) En annan lätt urskiljbar enhet är (det enligt någon kodstandard kodade) tecknet, och det är möjligt att vissa sorters lingvistisk informationsutvinning (statistiskt baserad induktion av distributionsklasser) skulle underlättas av att man startade från teckennivån och arbetade sig uppåt i den lingvistiska hierarkin. Detta är något som vi ämnar utforska vidare inom projektet.

Litteratur

Borin, Lars 1998. Linguistics isn't always the answer: Word comparison in computational linguistics. Under utgivning i *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Copenhagen, January 28-29, 1998. University of Copenhagen, Centre for Language Technology.

Prütz, Klas 1997. Sammanställning av en träningskorpus på svenska för träning av ett automatiskt ordklasstaggningssystem. Uppsala universitet, Institutionen för lingvistik.

Sågvall Hein, Anna 1997. Using parallel corpora in multi-lingual lexical acquisition. Under utgivning i H. Kalverkämper och B. Svane (utg.),

Übersetzen und Dolmetschen. Forschungsstand und Perspektive. Translation and Interpreting. State and Perspectives. Proceedings from the Humboldt-Stockholm Symposium, Stockholm University, January 30-31, 1997.

Tiedemann, Jörg 1997. Automatical lexicon extraction from aligned bilingual corpora. Diplomarbete i datavetenskap vid Otto-von-Guericke-Universität Magdeburg.

Tiedemann, Jörg 1998. Extraction of translation equivalents from parallel corpora. Under utgivning i *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Copenhagen, January 28-29, 1998. University of Copenhagen, Centre for Language Technology.

Tjong Kim Sang, Erik F. 1996a. Converting the Scania Framemaker documents to TEI SGML. Uppsala universitet, Institutionen för lingvistik.

Tjong Kim Sang, Erik F. 1996b. Aligning the Scania corpus. Föredrag vid översättningsprogrammets andra arbetsseminarium, Utö, 12–13 juni 1996. Uppsala universitet, Institutionen för lingvistik.